A Collection of Features for Semantic Graphs

Tina Eliassi-Rad, Imola K Fodor, Brian Gallagher

September 19, 2006

1 Introduction

Semantic graphs are commonly used to represent data from one or more data sources. Such graphs extend traditional graphs by imposing *types* on both nodes and links. This type information defines permissible links among specified nodes and can be represented as a graph commonly referred to as an *ontology* or *schema* graph. Figure 1 depicts an ontology graph for data from National Association of Securities Dealers. Each node type and link type may also have a list of attributes.

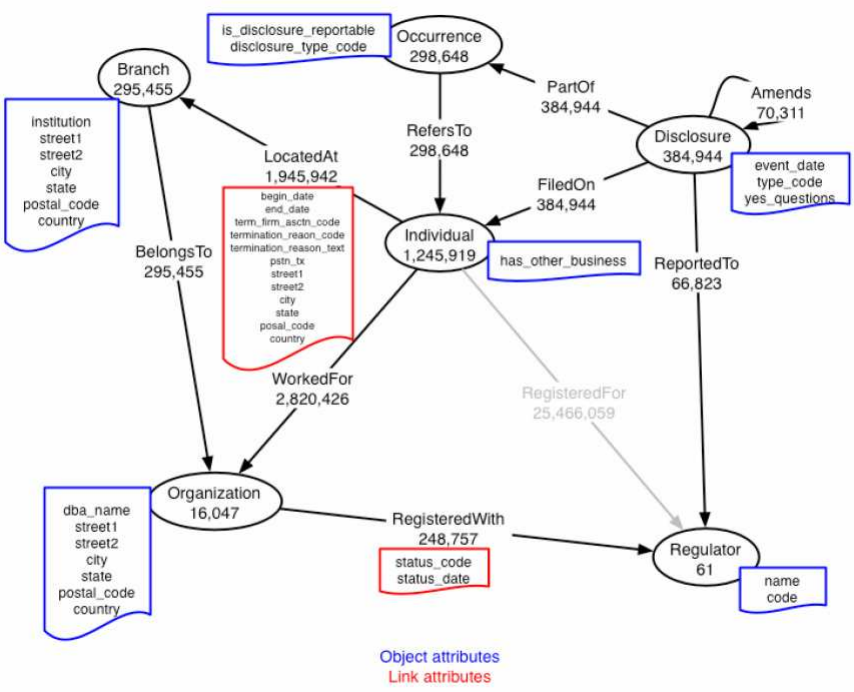


Figure 1: An Ontology Graph for Data from National Association of Securities Dealers (NASD). Courtesy of David Jensen, University of Massachusetts.

To capture the increased complexity of semantic graphs, concepts derived for standard graphs have to be extended. This document explains briefly features commonly used to characterize graphs, and their extensions to semantic graphs [1].

This document is divided into two sections. Section 2 contains the feature descriptions for static graphs. Section 3 extends the features for semantic graphs that vary over time.

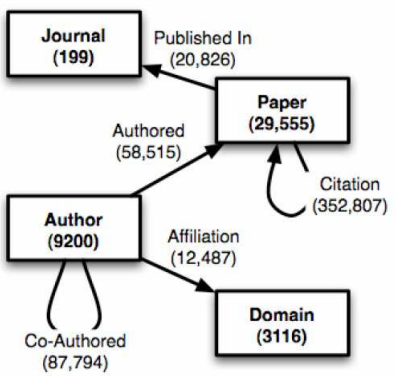


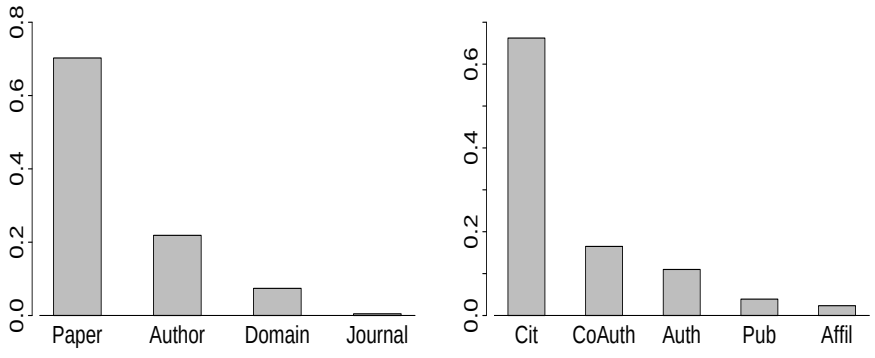
Figure 2: An ontology graph for a collaboration network [4]. The size of the semantic graph is given by the number of nodes $N = 42070 = 199 + 29555 + 9200 + 3116$ and the number of links $M = 532429 = 20826 + 352807 + 58515 + 87794 + 12487$. The size of the ontology graph is given by the number of node types $n = 4 = \{$Journal, Paper, Author, Domain$\}$ and the number of link types $m = 5 = \{$Published In, Citation, Authored, Co-Authored, Affiliation$\}$.

2 Features for semantic graphs

1. Size of a semantic graph
   (a) Number of nodes or vertices: N
   (b) Number of links or edges: M
2. Size of an ontology graph
   (a) Number of node types: n
   (b) Number of links types: m

   Fig. 2 illustrates the first two sets of features, using the collaboration network from [4].
3. Type distribution
   (a) Node type distribution: {the number of nodes of type n_i}/N, for each $i = 1, \ldots, n$
   (b) Link type distribution: {the number of links of type m_i}/M, for each $i = 1, \ldots, m$

   Fig. 3 presents the type distributions for the example in Fig. 2.
4. Degree of a node: node degree is generally defined based on the number of neighbors of a node, that is, on the number of nodes adjacent to it. An alternative definition is based on the number of edges incident to the respective node. While the two definitions are equivalent for most graphs, their results are different for multigraphs and graphs with self-loops.
   (a) In-degree of node i: the number of adjacent nodes with an edge into the node (alternatively, the number of incident edges into the node)
   (b) Out-degree of node i: the number of adjacent nodes with an edge out of the node (alternatively, the number of incident edges out of the node)
   (c) Degree of node i: k_i is the sum of the in-degree and the out-degree of node i
5. Average node degree or connectivity: $\sum_{i=1}^{N} k_i/N$. For undirected graphs with no self-loops, the average node degree is $2M/N$.



(a) (b)

Figure 3: Type distributions for the example in Fig. 2 for (a) nodes and (b) links. The value of the node type distribution for `Paper` is 29555/42070 ≈ 0.70. The value of the link type distribution for `Citation` is 352807/532429 ≈ 0.66.

6. Degree distribution: let $d_1, \ldots, d_D$ denote the D distinct values of the N node degrees. The degree distribution is the {counts of the nodes with degree d_i}/N, for $i = 1, \ldots, D$.

   The degree distribution provides a measure of structural homogeneity of the graph.

7. Degree of a node based on its type

   (a) Number of neighbors of type β of a node i of type α: $k_{\alpha\beta}(i)$

   (b) Number of neighbors of node i of type α: $k_\alpha(i) = \sum_\beta k_{\alpha\beta}(i)$

8. Average number of neighbors per type is a feature that provides a way to analyze connectivity based on semantic types

   (a) Average degree of a node of type α: $\overline{k_\alpha} = \frac{1}{N_\alpha} \sum_i k_\alpha(i)$, where type(i)=$\alpha$ and N_α is the number of nodes of type α in the semantic graph

   (b) Average number of neighbors per type, re-scaled to compare different types: $\mu_\alpha = \overline{k_\alpha}/k^0_\alpha$, where k^0_α is the number of node types to which a node of type α can connect

9. Standard deviation of the number of neighbors per type:

$$\sigma^k_\alpha = \frac{\sqrt{\overline{k^2_\alpha} - (\overline{k_\alpha})^2}}{k^0_\alpha}, \text{ where } \overline{k^2_\alpha} = \frac{1}{N_\alpha} \sum_{i:type(i)=\alpha} k^2_\alpha(i) \tag{1}$$

   The quantities μ_α and σ^k_α provide the expected number of connections of a node of a given type, and its dispersion, respectively.

10. Correlation between node degrees at either ends of an edge, or assortative mixing [5]: Let j_i and k_i denote the degrees of the nodes at the end of the ith edge. M continues to represent the number of links in the semantic graph. Then,

$$r = \frac{M^{-1} \sum_i^M j_i k_i - [M^{-1} \sum_i^M (\frac{j_i + k_i}{2})]^2}{M^{-1} \sum_i^M (\frac{j_i^2 + k_i^2}{2}) - [M^{-1} \sum_i^M (\frac{j_i + k_i}{2})]^2}. \tag{2}$$

   This concept extends the Pearson correlation coefficient between the node degrees for graphs. The values for r are in the interval [-1, 1], and indicate whether high-degree nodes tend to be connected to

other high-degree nodes. Social networks (such as co-authorship and collaboration networks) are often assortatively mixed ($r > 0$), while technological and biological networks (such as the Internet, protein interactions, neural networks) tend to be dissortative ($r < 0$).

11. Disparity of connected types, or correlation between different types: Some node types have connections to many other node types, even when these nodes are not connected to many other nodes.

    (a) First calculate relevant node types Y_2, such that a small value indicates a large number of relevant types, and large value indicates the dominance of a few types

    $$Y_2(i;\alpha) = \sum_{\beta} \left[\frac{k_{\alpha\beta}(i)}{k_{\alpha}(i)} \right]^2 \tag{3}$$

    (b) Next calculate average and dispersion over all nodes of the same type

    $$\overline{Y}_2(\alpha) = \frac{1}{N_\alpha} \sum_{i:type(i)=\alpha} Y_2(i;\alpha);\ \ \overline{Y}_2^2(\alpha) = \frac{1}{N_\alpha} \sum_{i:type(i)=\alpha} Y_2^2(i;\alpha);\ \ \sigma_\alpha^Y = \sqrt{\overline{Y}_2^2(\alpha) - (\overline{Y}_2(\alpha))^2} \tag{4}$$

    (c) Finally, normalize both the average and the dispersion by population

    $$Y_2^r = \sum_{\beta \in \nu(\alpha)} \left[\frac{N_\beta}{N} \right]^2 \text{ and } R(\alpha) = \frac{\overline{Y}_2(\alpha)}{Y_2^r} \text{ and } \sigma_\alpha^R = \frac{\sigma_\alpha^Y}{Y_2^r} \tag{5}$$

    $\nu(\alpha)$ is the set of node types that can be connected to a given node α as dictated by the ontology. N_β is the number of nodes of type β in the semantic graph.

    High disparity node types are often not relevant for path finding (i.e. search). Semantically similar nodes tend to have similar values of average number of neighbors per type, and similar values of disparity.

12. Average path length is the average of the values in D, where D denotes the matrix of the N^2 shortest distances d_{ij} between the node pairs in the graph, $i, j = 1, \ldots, N$:

    $$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{pmatrix}. \tag{6}$$

    For directed graphs, D is not necessarily symmetric.

    For semantic graphs, type-dependent average path lengths can be calculated, in which case D is the set of shortest distances between node pairs where either the source or the destination node is of a particular type.

13. Diameter for all nodes and per type: the maximum value of the distances d_{ij} in the distance matrix D in Eq. (6).

    In a semantic graph, type-dependent diameters are defined similarly to the type-dependent average path length in feature 12.

14. Breadth-first search (BFS) level distance from a node: Given a node, the breadth-first level distances to the remaining N-1 nodes, where the BFS level distance is defined as the number of edges between start and end nodes along the shortest path.

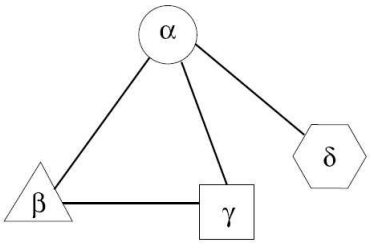


Figure 4: An example ontology for which neighbors of α of type δ can never be connected to neighbors of type β or γ. Thus, when measuring the clustering coefficient for α, δ should not be penalized for not being connected to β or γ.

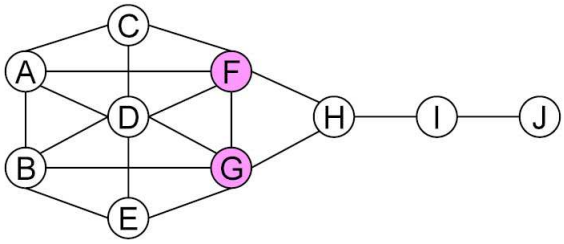


Figure 5: An example graph illustrating the closeness centrality concept. The closeness scores in decreasing order are: F: 0.643, G: 0.643, D: 0.600, H: 0.600, A: 0.529, B: 0.529, C: 0.500, E: 0.500, I: 0.429, J: 0.310.

15. Clustering coefficient for node i:

$$C(i) = \frac{E_i}{k_i(k_i - 1)}, \tag{7}$$

where k_i is the number of neighbors of node i and E_i is the number of edges between the k_i nodes.

The average over all nodes gives the clustering coefficient of the graph.

Within social networks, the clustering coefficient captures the common belief that a friend of a friend is also a friend. Graphs that exhibit the small-world property (i.e. most pairs of nodes are connected by a short path through the graph) have high clustering coefficient and low diameter (and, therefore, short average path length).

16. Type-dependent clustering coefficient for node i of type α in a semantic graph:

$$C(i;\alpha) = \frac{E_i}{E(i;\alpha)}, \tag{8}$$

where $E(i;\alpha)$ denotes the maximum number of links allowed by the ontology. The simple example in Fig. 4 illustrates this concept.

17. Closeness centrality of node i: the ratio of 1 over its average geodesic (i.e., shortest) distance to all other nodes in the graph – namely, $cls_i = (\frac{1}{N-1}\sum_{t\neq i\in V} d_G(i,t))^{-1}$, where $d_G(i,t)$ is the geodesic distance between node i and node t. A node with high closeness centrality can access all the nodes in the graph more quickly than other nodes. Such a node has the shortest paths to the other nodes, and can easily monitor the information flow in the graph. For example, in the graph presented in Fig. 5, nodes F and G have the highest closeness scores.

Extension to semantic graphs is provided by the ratio of 1 over the its average distance to all nodes of type α in the graph.

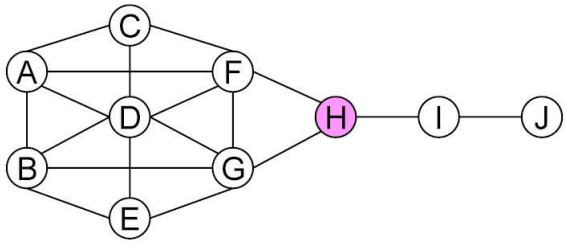


Figure 6: An example graph illustrating the betweenness centrality concept. The betweenness scores are: H: 0.389, F: 0.231, G: 0.231, I: 0.222, D: 0.102, A: 0.023, B: 0.023, C: 0.000, E: 0.000, J: 0.000.

18. Betweenness centrality

    (a) Node betweenness: $bet_i = \frac{1}{\frac{1}{2}N(N-1)} \sum_{s \neq i \neq t \in V} \frac{g_i(s,t)}{N_{st}}$, where $g_i(s,t)$ is the number of geodesic (i.e. shortest) paths from node s to node t that pass through node i. N_{st} is the total number of geodesic paths from s to t. V is the set of nodes in the graph and N is the total number of nodes (i.e., $N = |V|$).

    (b) link betweenness: $bet_{i \to j} = \frac{1}{\frac{1}{2}N(N-1)} \sum_{s \neq i \neq j \neq t \in V} \frac{g_{i \to j}(s,t)}{N_{st}}$, where $g_{i \to j}(s,t)$ is the number of geodesic (i.e. shortest) paths from node s to node t that traverse the link connecting i to j.

    A node or a link with a high betweenness has great influence over what flows in the network. For example, in the graph presented in Fig. 6, node H has the highest betweenness centrality value.

    For semantic graphs, these concepts are extended to the fraction of shortest paths between node pairs that pass through a node i of type α or an edge j of type θ.

19. Relevance of node i: defined as the clustering coefficient $C(i)$ if the links of node i are all of the same type (i.e. count the links involving pairs of neighbors of node i). Otherwise, given by a matrix $M(t_1, t_2)$ that counts the number of links between pairs of neighbors a and b, where node a is linked to node i via type t_1 and node b is linked to node i via type t_2.

    Small entries in $M(t_1, t_2)$ correspond to pairs of link types associated with node i that should not be traversed in path-finding (i.e. search).

20. Relevance of a link between nodes a and b: quantifies the relevance of the relationship between a and b by counting the proportion of neighbors they share

$$S(a,b) = \frac{|I(a,b)|}{|U(a,b)|}, \tag{9}$$

    where I(a, b) = {node $w | w$ is linked to a and b, $w \neq a, w \neq b$} and
    U(a, b) = {node $w | w$ is linked to a or b, $w \neq a, w \neq b$} = degree(a) + degree(b) − I(a, b).

    Large values of $S(a,b)$, $0 \leq S(a,b) \leq 1$, indicate strong relationships between nodes a and b, with a high proportion of common neighbors. Fig. 7 displays an example.

21. Connected graph: graph G is connected if every node is reachable from every other node, that is, if every two nodes are connected by *at least* one link. Otherwise, G is disconnected.

22. Clique: G is a clique if each pair of nodes is connected by a link. A clique with N nodes has $N(N-1)/2$ undirected links.

    Note: in a multigraph (see definition 26) a clique may have more than one link connecting each pair of nodes. In addition, a multigraph may have $>= N(N-1)/2$ links and not be a clique.

23. Component graph: a connected subgraph H of G is a component of G if H is not contained in any connected subgraph of G having more nodes or links than H.

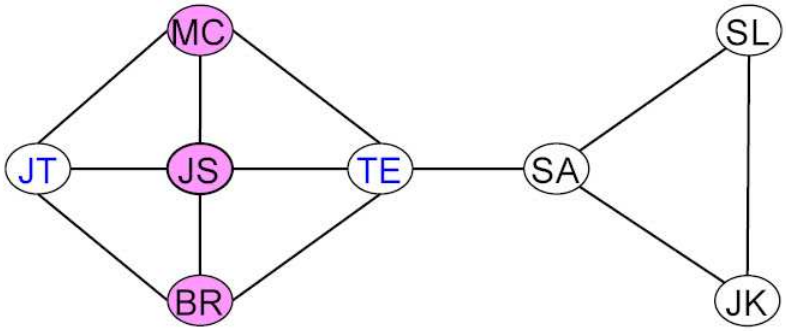


Figure 7: Example graph illustrating the relevance of a link between two nodes given in Eq. (9), with S(JT,TE)=0.75.

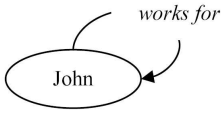


Figure 8: An example self-loop where a node connects to itself.

24. Largest (or strongly) connected component (SCC): the SCC of G is the largest subgraph in G in which every node is reachable from every other node. Given a graph $G = \{V, E\}$, its strongly connected component is a subgraph $SCC = \{V', E'\}$ that (i) is connected and (ii) for all nodes u such that $u \in V$ and $u \notin V'$, there is no node $v \in V'$ for which u has a link to v.

25. Self-loop: G has self-loops when a node can have a link to itself. For example, self-employment in Fig. 8 would constitute a self-loop in a graph.

    Self-loops are a special case of graphs that contain cycles. Depending on the application, node degree can be calculated with or without including the self-loops.

26. Multigraph: G is a multigraph if it allows multiple (parallel) links between two nodes. Figure 9 depicts a simple multigraph. A simple way to measure the prevalence of multiple links is to first collapse the multiple links between each pair of nodes into a single link. Let G' denote the resulting new graph, and M and M' denote the number of edges in the original and new graphs, respectively. Then, $\frac{M-M'}{M'}$ measures the

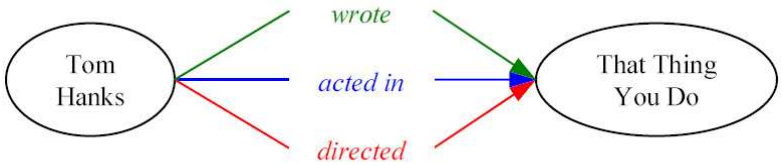


Figure 9: Example multigraph where multiple links – wrote, acted in, and directed – connect the actor node Tom Hanks to the movie node That Thing You Do.

27. Edge redundancy measures the graph's robustness to random disconnections, and can be measured by the average number of randomly selected links that must be removed to break the graph into disconnected components. Details on the random edge removal and subsequent calculation of the redundancy are found in [3].

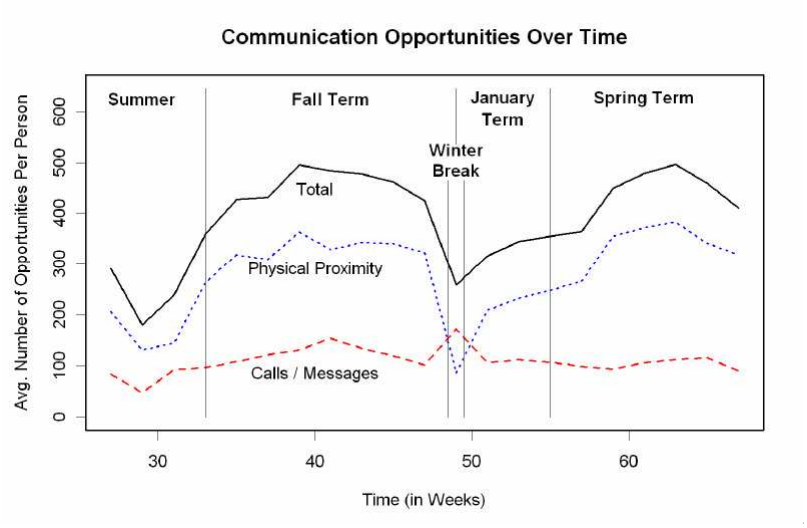


Figure 10: The average number of communication opportunities per person, calculated over time, for the MIT Reality Mining data [2].

28. Other features such as strength of association [3], graph spectra (eigenvalues of the adjacency matrix), PageRank, and HITS will be considered as the project progresses.

3 Extensions to dynamic graphs

The features presented in Section 2 can easily be extended to semantic graphs that vary over time. Given a start time, an end time, and an interval length (i.e. size of time step), the graph can be divided into subgraphs that exist in the time intervals defined by these three parameters. The selected features are then calculated for each of the subgraphs, resulting in a time series of features. Monitoring the evolution of features through time can uncover temporal signatures and anomalies in the graph.

Fig. 10 was generated by calculating graph features over time on the MIT Reality Mining Data set [2]. It shows the number of communication opportunities (i.e., calls and voice and text messages between phones and physical proximity between Bluetooth devices) per person over time. These values were calculated using feature 2b (i.e., counting the number of links) on the communication links between cell phones and the proximity links between Bluetooth devices over time. The plot shows that the number of physical meetings between participants is high during each term and falls off during breaks. In contrast, the number of calls remains relatively constant over time and peaks during winter break, presumably because participants have more time to call and may want to keep in touch with friends from school while physical meetings are impossible.

Acknowledgments

This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

References

[1] M Barthélemy, E Chow, and T Eliassi-Rad. Knowledge representation issues in semantic graphs for relationship detection. In *Papers from the 2005 AAAI Spring Symposium - AI Technologies for Homeland Security, Stanford, CA*, pages 91–8, 2005.

[2] N Eagle and A Pentland. Reality mining: sensing complex social systems. *J of Personal and Ubiquitous Computing*, 10(4):255–68, 2006. http://reality.media.mit.edu/pdfs/realitymining.pdf.

[3] TL Hickling and WG Hanley. Methodologies and metrics for assessing the strength of relationships between entities within semantic graphs. Technical Report UCRL-TR-216074, Lawrence Livermore National Laboratory, 2005.

[4] A McGovern, L Friedland, M Hay, B Gallagher, A Fast, J Neville, and D Jensen. Exploiting relational structure to understand publication patterns in high-energy physics. *SIGKDD Explorations*, 5(2):165–73, 2003.

[5] MEJ Newman. Assortative mixing in networks. *Phys Rev Lett*, 89(208701):1–4, November 2002.